The weighted values of solar evaporation's environment factors obtained by machine learning

Yunpeng Wang [a,b,#], Guilong Peng [a, b, #], Swellam W. Sharshir [b,c], AbdAllah W. Kandeal [a, b,], Nuo Yang [a,b] *

[a] State Key Laboratory of Coal Combustion, and [b] School of Energy and Power Engineering, Huazhong University of Science and Technology, Wuhan 430074, China.

[c] Mechanical Engineering Department, Faculty of Engineering, Kafrelsheikh University, Kafrelsheikh 33516, Egypt.

[#] YW and GP contributed equally on this work.

*Corresponding authors: NY (nuo@hust.edu.cn)

Abstract

Enhancing the efficiency of solar evaporation is important for solar stills. In this study, the weighted values of environment factors (descriptors) on the efficiency of solar evaporation are obtained by using a machine learning algorithm, random forest. To verify the advancement between random forest and mathematical data analysis, two traditional methods, pair wise plots and Pearson correlation analysis, are conducted for comparison. Experimental data are obtained from around 100 articles since 2014. The results indicated that traditional methods failed at obtaining reasonable weighted values, while random forest is competent. It is found that thermal design is the most significant descriptors to obtain a high efficiency. The lack of complete dataset is the main challenge for more in-depth and comprehensive analysis. This work may promote the studies on solar evaporation and solar stills.

Key words: Solar still; Solar evaporation; Machine learning; Environment factors

1. Introduction

With the population growth and the development of industrial activities and agricultural progress, the shortage of fresh water resources is becoming one of the catastrophic problems that the world faces. Given that sea water account for 97% plant's water resources, it is desirable to develop technologies for sea water desalination. Many effective methods of desalination have been proposed in the past, like multistage flashing[1,2], reverse osmosis, multi-effect distillation and vapor compression[3,4] and so on.

Compared to other methods, solar still attracts more and more interest due to its eco-friendly, simple construction and maintenance, low installation cost and long life operation[5,6]. Solar evaporation is one of the crucial process in solar still. Thereby, during past decades, many methods have been proposed to achieve high efficiency of solar evaporation, such as using nanofluid[7,8], cotton cloth[9], sponge and charcoal[10].

However, the evaporation efficiency is affected by many factors, such as materials type, thermal design, ambient temperature, solar intensity and so forth. Therefore, it is interesting to show the importance or weighting of different factors. Empirically, several important factors can be picked up. However, it is difficult to quantify the importance of each descriptor, thus few works discussed this point in solar evaporation field. On the other hand, quantitative analysis of the descriptor importance is a widespread scientific problem. Such as in the field of chemistry, machine learning technology[11-14] was used to measure the descriptor importance[15,16] for polymer chain angle, Random forest for example[17]. It is found that machine learning can accurately measure the relationship between different factors (descriptors) and their influence on the target value, which present a meaningful first step towards the high-throughput screening of polymer chemistry to identify compositions with desirable bulk properties.

In the current study, three methods are used and compared for obtaining the weighted values of each descriptor in solar evaporation. Including two traditional data science methods, i.e. pair wise plots (PWP)[18] and Pearson correlation analysis (PCA)[19], and a machine learning algorithm, random forest (RF)[20]. The weighted values also called descriptor importance. Firstly, the traditional data science methods are used. Pair wise plots and Pearson correlation analysis are used for measuring correlations between pairwise descriptors. Then, random forest is conducted for measuring the importance between evaporation efficiency and descriptors. The results of descriptor importance can eventually help scientists to design high efficiency system of solar evaporation.

2. Methodology

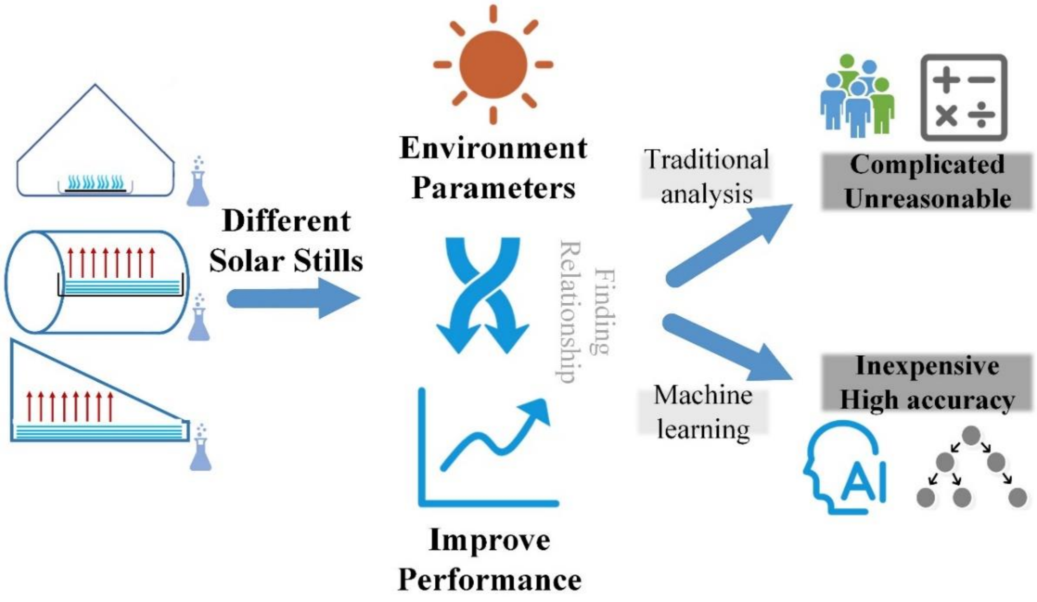

Fig. 1 main flowchart of current study.

The main flowchart of current study is showed in Fig. 1, experimental data used in analysis are collected from around 100 articles since 2014 (Details in Supporting Materials).

Table 1 The details of data representation

Descriptors	Classification	Labels	Number of samples
Solar intensity	1 Kw	0	51
	1-10 Kw	1	25
	>10 Kw	2	10
Thermal design	3D interface	0	53
	2D\1D interface	1	21
	Volumetric	2	12
Surface diameter	<3 cm	0	30
	3-4 cm	1	29
	>4 cm	2	29
Absorptivity	<0.95	0	23
	0.95	1	31
	>0.95	2	34
T_{amb}	<24℃	0	21
	24-25℃	1	40
	>25℃	2	27
T_{vapor}	<50℃	0	37
	50-70℃	1	31
	>70℃	2	20
Efficiency	<75 %	0	28
	75 %~85 %	1	31
	>85%	2	28

For the collected dataset, $\mathrm{M} = \{\mathrm{X}, \mathrm{y}\}_{1:t}$, y is the objective value (energy efficiency of evaporation), X are input descriptors (descriptors) corresponding to y, such as the solar intensity, thermal design, surface diameter, absorptivity, T_{amb} (temperature of ambient) and $T_{interface}$ (temperature of interface). Due to the lack of details in original articles, there are some missing data of surface diameter, absorptivity, T_{amb} and $T_{interface}$. The method of Mean Completer is used for filling missing data. Each descriptor is divided into three labels. The detailed distribution of dataset M is list in Table 1.

Three methods, including pair wise plots (PWP), Pearson correlation analysis (PCA), and machine learning algorithm, i.e. random forest (RF) are studied in this work. The application of three methods can be summarized as follows:

2.1 Pearson correlation analysis

PCA is usually being adopted to quantify the correlation between two different descriptors. For descriptors X_1 and X_2 in M, such as solar intensity and solar absorptivity, PCA can be calculated as Eq.1:

$$\rho = \frac{\sum(X_1 - \overline{X_1})(X_2 - \overline{X_2})}{[\sum(X_1 - \overline{X_1})^2 \sum(X_2 - \overline{X_2})^2]^{1/2}} \tag{1}$$

The values of PCA are dimensionless and distribute from -1 to 1. The closer to 1 or -1, the stronger correlation between two descriptors. The value 0 suggests no correlation. If the value is positive, then a positive correlation exists, else a negative correlation exists.

2.2 Pair wise plots

Pair wise plots can draw scatterplots for descriptor correlation and histograms for univariate distributions intuitively. As presented in Fig. 3, the subfigures on diagonal represents the data distribution trend of a particular descriptor. The other subfigures can help us visibly and qualitatively observe the relationship between each pair of descriptors. If a rising or falling trend is formed on the diagonal, the corresponding pair of descriptors have strong correlation. Otherwise, the two are not correlated.

2.3 Random forest

Random forest is a typical ensemble method, which combines multiple decision trees into one model to improve the performance. It is widely applied in many scientific and engineering fields, such as statistics, materials and biology[21,22]. The main step of RF shown in Fig. 2 can be expressed as

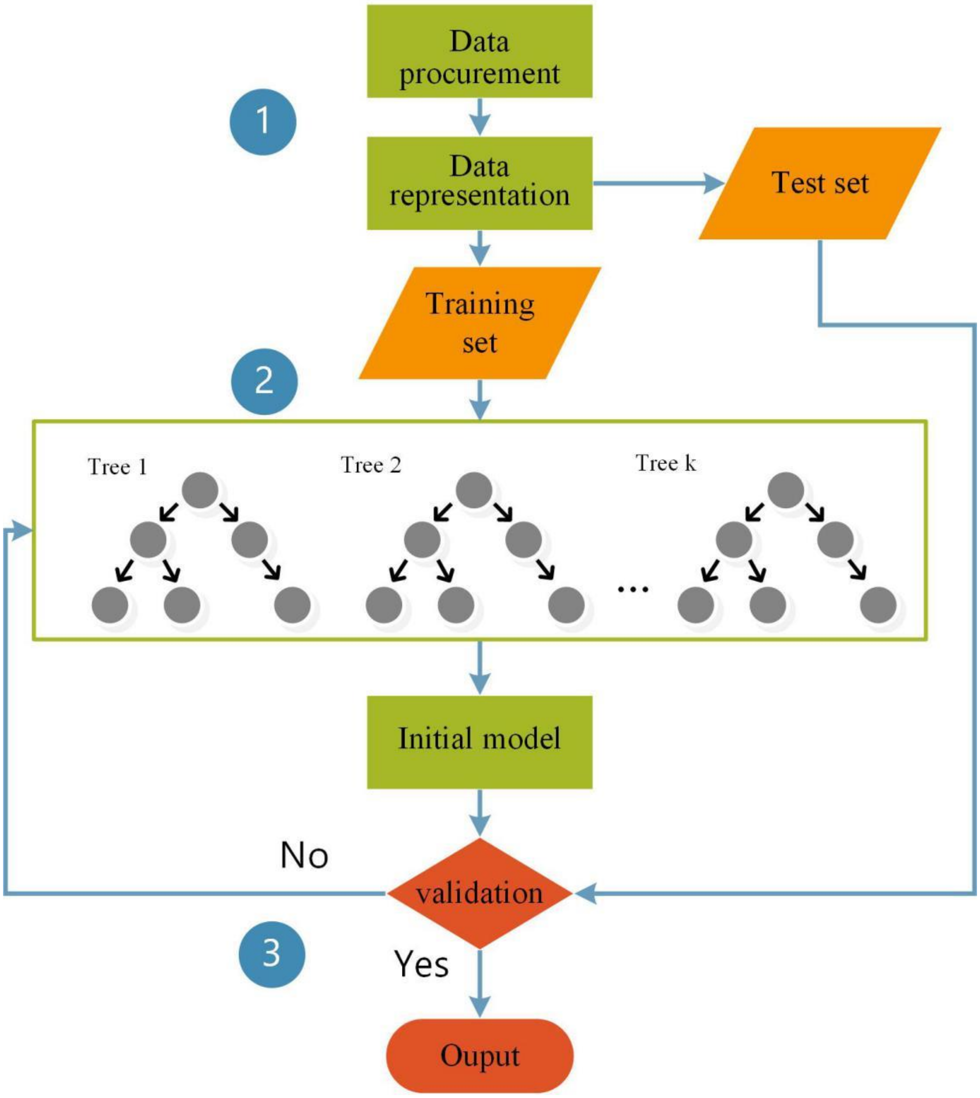

Fig.2 Schematics of applying the random forest in studying the importance of different descriptors.

(1) Data preprocessing

In the present study, data representation is performed for converting data into symbols which can be read by computers. For instance, three types of thermal design, 3D interface, 2D/1D interface, and volumetric, are represented as 0, 1, and 2 (details in Table 1). Finally, the dataset is divided into training set TA and test set TE according to a certain ratio.

(2) Model construction

Based on a processed dataset TA, the bootstrap resampling method[23] is used to randomly generate K sets of data. Then, K decision trees will be grown. For example, in calculating Fig. 5a, each dataset includes three descriptors (thermal design, absorptivity and random descriptor) and the label (efficiency). In each node of a decision tree, the node will split the dataset into two parts according to the value of a chosen descriptor. After traverse all descriptors, the final node will be the label (efficiency) of this data. The prediction of the model is voted by K decision trees.

(3) Model validation

Test dataset TE, which is not trained in model construction, is used for judging the accuracy of the model. If the accuracy of TA is much higher than the accuracy of TE (e.g. 0.9 for TA and 0.5 for TE), the model is considered as overfitting. If the accuracy of TE and TA are too low (e.g. 0.5 for TA and 0.5 for TE), the model is considered as underfitting. Both overfitting and underfitting are unacceptable, in which cases, the model needs to be retrained. If the accuracy of TA and TE is high enough and the accuracy of TE is similar as TA, the model is consider as trained well.

3. Result and discussion

As a starting point, two traditional data science method, Pair wise plots and Pearson correlation analysis, are performed for finding the weighted values, i.e. the descriptor importance. The results are displayed in Fig. 3 and Fig. 4. In addition to the mathematic data analysis, well-established machine learning algorithms are used to extract the relationship between descriptors and the target property in materials informatics. Subsequently, Fig. 5 showed the results that calculated by machine learning algorithm, Random forest.

3.1 Result of traditional data analysis

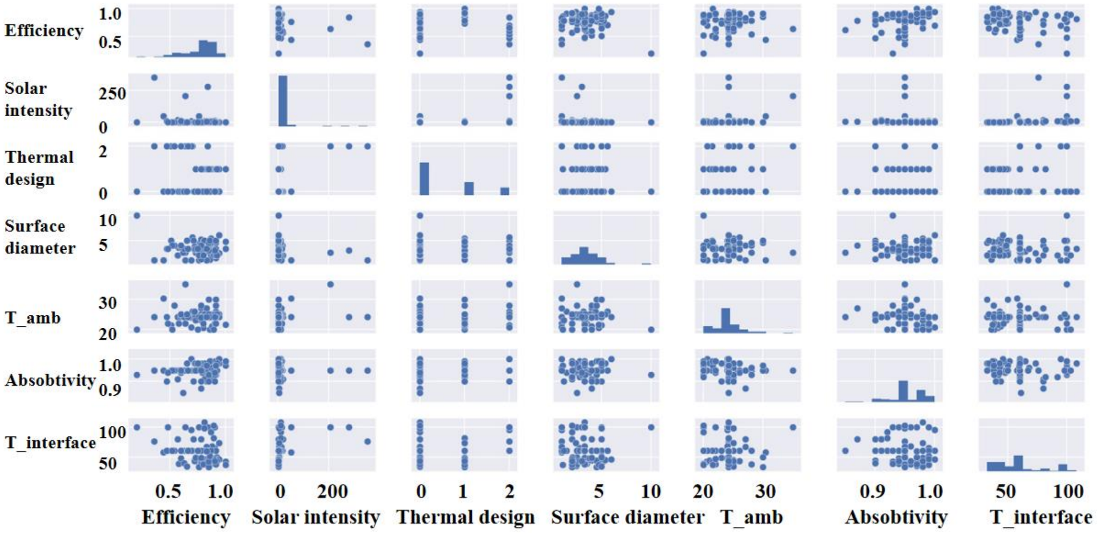

Fig. 3 Pair wise plots of dataset

Fig. 3 shows the pair wise plots between different descriptors (environment descriptors) and efficiency of solar evaporation. As can be seen from the top row of plots in Fig. 3, there is no obvious linear relation between efficiency and other descriptors. Moreover, it is clear that the dataset is discrete and not evenly distributed. Most of solar intensity data is around 1 kW and the thermal design is set as discrete numbers. Since PWP is the method which focus on the pure mathematic map in dataset, it is reasonable that PWP can't measure the descriptor importance based on defective dataset.

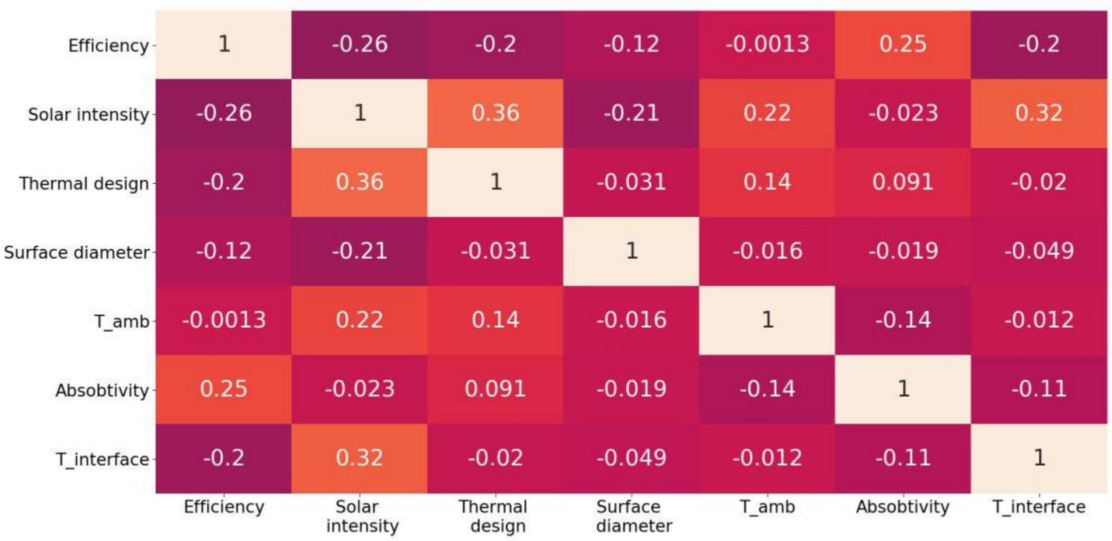

Fig. 4 Map of Pearson correlation analysis

The values of PCA are displayed in Fig. 4. As can be seen from all investigated descriptors, all absolute descriptors' values are below 0.3, which means all descriptors have weak correlation with the efficiency. Therefore, similar to PWP, it is unlikely to draw reasonable results of descriptor importance based on the PCA values.

3.2 Result of machine learning algorithms

Fig.5a-5d showed the descriptor importance quantified by RF of using 2, 3,4, and 6 selected descriptors, respectively. The results indicated that thermal design is the most important descriptor in solar evaporation, among all chosen descriptors. The importance of thermal design is at least 2 times higher than other descriptors. This result showed that optimizing the heat transfer process in solar evaporation system is essential for enhancing the efficiency of solar evaporation. This result is reasonable, because of that when the thermal design is poor, only a small part of solar energy is used for evaporation. For example, in a traditional solar evaporation, some heat is used for heating the bulk water, instead of for promoting evaporation[24]. Therefore, the efficiency of solar evaporation of volumetric system is lower than that of the interface system. Besides, Fig.5 shows that solar intensity is an unimportant descriptor. This is because that, with the optimized thermal and material design, high efficiency can be obtained for no matter high or low solar intensity as reported in many works[25-27]. Therefore, solar intensity was not important and similar to a random descriptor. Herein, the random descriptor is a set of random data which has no relationship to the energy efficiency and is used as a benchmark.

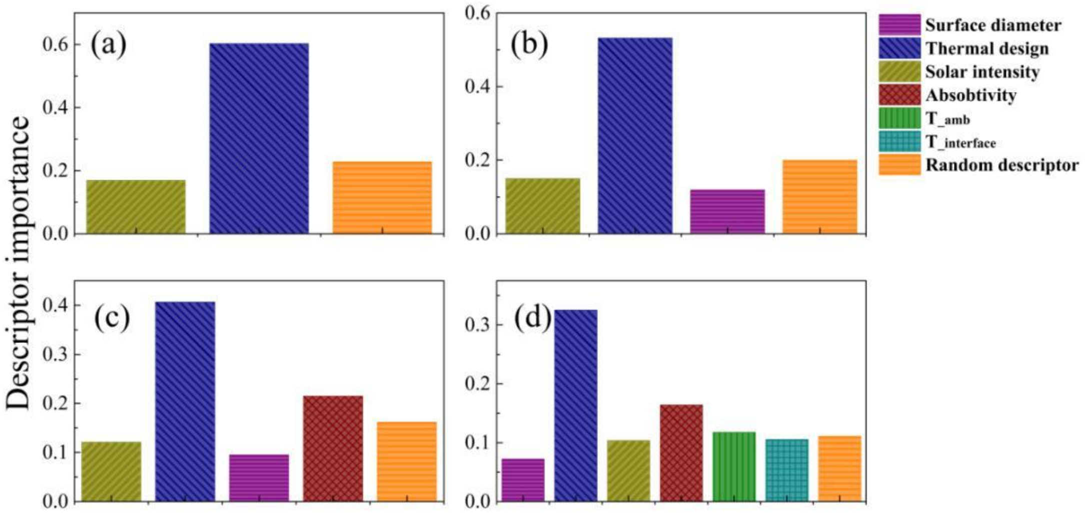

Fig.5 The result of descriptor importance by using RF. The sum of values of descriptor importance equals 1. (a)-(d) are the results of using 2, 3,4, and 6 different descriptors, respectively. Surface diameter is the diameter (length) of the evaporation surface.

Meanwhile, the descriptor importance of solar absorptivity is much lower than expectation as shown in Fig.5. Higher absorptivity enables more available energy for evaporation and will affect the efficiency a lot. The reason maybe that almost all reported works picked up materials with very high absorptivity (>90%), which makes the dataset cannot reach ergodicity. Hence, its importance is underestimated in the calculation. Besides, the temperature of ambient (T_{amb}) and evaporation interface ($T_{interface}$) are insignificant, which might due to the small difference of T_{amb} and $T_{interface}$ between most of works. However, temperature is actual a very important descriptor in natural convection based evaporation process[28]. Therefore, to capture the real importance of temperature, more works should be done at different ambient and interface temperature.

It can be concluded that a more accurate calculation by machine learning requires more complete data. Compared to other fields such as materials[29,30], they have a complete database of physical properties and theoretical calculation methods. However, in the current study, the shortage of reported data is a hard problem because authors do not provide exact values of some descriptors. For example, values of ambient temperature, the diameter of evaporation surface, absorptivity, and the temperature of evaporation interface are missing in some papers. Therefore, to obtain a more accurate result of machine learning, authors should provide complete dataset of experimental descriptors in their future works. On the other side, some other potential important descriptors on material design, such as thermal conductivity, contact angle, specific area, porosity, characteristic size, functional group and so forth, are not included and calculated by RF in the current stage, because the detailed properties of materials are not offered in most of papers. It is worth to be noticed that a full dataset of descriptors in research reports will help to push the field forward.

3.3 Effect of the size of dataset

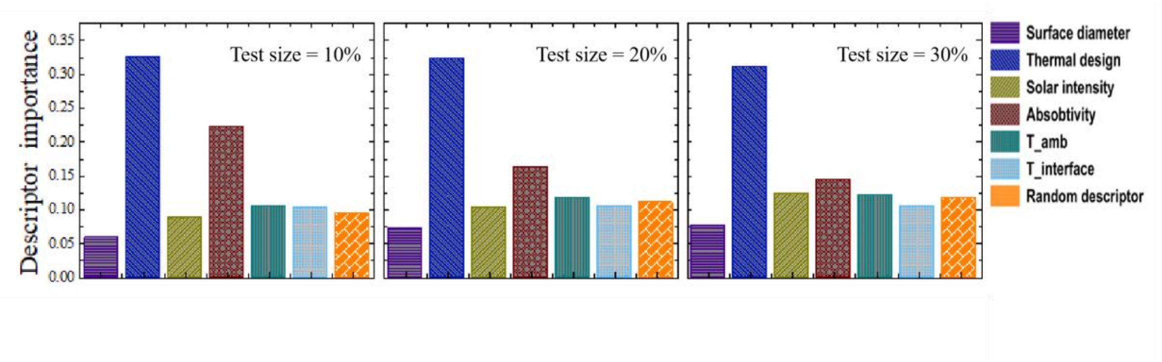

Fig. 6 The results of descriptor importance correspond to three combinations of train/test set.

As mentioned in other researches, the quality of dataset determines the reliability of machine learning algorithms. In order to avoid overfitting and further quantify the effect of dataset in the current study, the initial dataset was separated into three combinations of train/test set, i.e. 70%/30%, 80%/20% and 90%/10%, respectively. The results are summarized in Fig.6. As can be seen, with the decrease of test set, there is no obvious difference between different models. Thermal design is the most important descriptor in all cases. Absorptivity is the second important descriptor. Other descriptors were not important and similar to a random descriptor. It turns out that the result of descriptor importance depends on the physical mechanism rather than the dataset. This indicates that the current dataset is able to get a convergent solution.

4. Conclusion

In conclusion, the importance of factors on efficiency of solar evaporation are analyzed by pair wise plots, Pearson correlation analysis, and random forest. Experimental data used in analysis are collected from around 100 articles. The results indicate that pair wise plots, Pearson correlation analysis can't measure the descriptor importance based on defective dataset. On the contrary, random forest can obtained reasonable results. The results by using random forest show that thermal design is the most important descriptor that determining the efficiency of solar evaporation. It can be concluded that machine learning is helpful to understand the importance of various descriptors quantitatively, which will help to push the solar still field forward.

Although machine learning obtained meaningful results, it should be emphasized that due to the limitation of the amount and quality of experimental data in published articles, the current analysis is more about qualitative results than quantitative results. It is expected that authors can provide more detailed data and standardized descriptors in future publication. This will promote the application of machine learning in studying solar still.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgment

The work was sponsored by National Key Research and Development Project of China (2018YFE0127800), China Postdoctoral Science Foundation (2020M682411), National Natural Science Foundation of China (51950410592), Fundamental Research Funds for the Central Universities (2019kfyRCPY045) and Graduate Innovation Funds for Huazhong University of Science and Technology (2020yjsCXCY067). The authors thank the National Supercomputing Center in Tianjin (NSCC-TJ) and China Scientific Computing Grid (ScGrid) for providing assistance in computations.

References

1 Al-Othman, A., Tawalbeh, M., El Haj Assad, M., Alkayyali, T. & Eisa, A. Novel multi-stage flash (MSF) desalination plant driven by parabolic trough collectors and a solar pond: A simulation study in UAE. *Desalination* **443**, 237-244, doi:https://doi.org/10.1016/j.desal.2018.06.005 (2018).

2 Shaaban, S. Performance optimization of an integrated solar combined cycle power plant equipped with a brine circulation MSF desalination unit. *Energy Conversion and Management* **198**, 111794, doi:https://doi.org/10.1016/j.enconman.2019.111794 (2019).

3 Farsi, A. & Dincer, I. Development and evaluation of an integrated MED/membrane desalination system. *Desalination* **463**, 55-68, doi:https://doi.org/10.1016/j.desal.2019.02.015 (2019).

4 Sadri, S., Ameri, M. & Haghighi Khoshkhoo, R. Multi-objective optimization of MED-TVC-RO hybrid desalination system based on the irreversibility concept. *Desalination* **402**, 97-108, doi:https://doi.org/10.1016/j.desal.2016.09.029 (2017).

5 Sharshir, S. W. *et al.* Improving the solar still performance by using thermal energy storage materials: A review of recent developments. *Desalination and water treatment* **165**, 1-15, doi:https://doi.org/10.5004/dwt.2019.24362 (2019).

6 Sharshir, S. W. *et al.* Augmentation of a pyramid solar still performance using evacuated tubes and nanofluid: Experimental approach. *Applied Thermal Engineering* **160**, 113997, doi:https://doi.org/10.1016/j.applthermaleng.2019.113997 (2019).

7 Trisaksri, V. & Wongwises, S. Critical review of heat transfer characteristics of nanofluids. *Renewable and Sustainable Energy Reviews* **11**, 512-523, doi:https://doi.org/10.1016/j.rser.2005.01.010 (2007).

8 Elsheikh, A. H., Sharshir, S. W., Mostafa, M. E., Essa, F. A. & Ahmed Ali, M. K. Applications of nanofluids in solar energy: A review of recent advances. *Renewable and Sustainable Energy Reviews* **82**, 3483-3502, doi:https://doi.org/10.1016/j.rser.2017.10.108 (2018).

9 Kalidasa Murugavel, K. & Srithar, K. Performance study on basin type double slope solar still with different wick materials and minimum mass of water. *Renewable Energy* **36**, 612-620, doi:https://doi.org/10.1016/j.renene.2010.08.009 (2011).

10 Abu-Hijleh, B. A. K. & Rababa'h, H. M. Experimental study of a solar still with sponge cubes in basin. *Energy Conversion and Management* **44**, 1411-1418, doi:https://doi.org/10.1016/S0196-8904(02)00162-0 (2003).

11 Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547-555, doi: https://doi.org/10.1038/s41586-018-0337-2 (2018).

12 Ouyang, Y. *et al.* Accuracy of Machine Learning Potential for Predictions of Multiple-Target Physical Properties. *Chinese Physics Letters* **37**, 126301, doi: https://doi.org/10.1088/0256-307x/37/12/126301 (2020).

13 Brochu, E., Cora, V. M. & De Freitas, N. J. a. p. a. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv preprint arXiv:1012.2599 (2010).

14 Wan, X. *et al.* Materials Discovery and Properties Prediction in Thermal Transport via Materials Informatics: A Mini Review. *Nano Letters* **19**, 3387-3395, doi: https://doi.org/10.1021/acs.nanolett.8b05196 (2019).

15 Ma, R., Huang, D., Zhang, T. & Luo, T. Determining influential descriptors for polymer chain conformation based on empirical force-fields and molecular dynamics simulations. *Chemical Physics Letters* **704**, 49-54, doi:https://doi.org/10.1016/j.cplett.2018.05.035 (2018).

16 Wang, Y. *et al.* Prediction of tubular solar still performance by machine learning integrated with Bayesian optimization algorithm. *Applied Thermal Engineering* **184**, 116233, doi:https://doi.org/10.1016/j.applthermaleng.2020.116233 (2021).

17 Ma, R., Liu, Z., Zhang, Q., Liu, Z. & Luo, T. Evaluating Polymer Representations via Quantifying Structure–Property Relationships. *Journal of Chemical Information and Modeling* **59**, 3110-3119, doi: https://doi.org/10.1021/acs.jcim.9b00358 (2019).

18 Chen, X., Wang, M. & Zhang, H. The use of classification trees for bioinformatics. *WIREs Data Mining and Knowledge Discovery* **1**, 55-63, doi:https://doi.org/10.1002/widm.14 (2011).

19 Calzetta, L. *et al.* Pharmacological treatments in asthma-affected horses: A pair-wise and network meta-analysis. *Equine Veterinary Journal* **49**, 710-717, doi:https://doi.org/10.1111/evj.12680 (2017).

20 Månsson, R. *et al.* Pearson Correlation Analysis of Microarray Data Allows for the Identification of Genetic Targets for Early B-cell Factor*[boxs]. *Journal of Biological Chemistry* **279**, 17905-17913, doi:https://doi.org/10.1074/jbc.M400589200 (2004).

21 Breiman, L. Random Forests. *Machine Learning* **45**, 5-32, doi: https://doi.org/10.1023/A:1010933404324 (2001).

22 Maltecca, C. *et al.* Predicting Growth and Carcass Traits in Swine Using Microbiome Data and Machine Learning Algorithms. *Scientific Reports* **9**, 6574, doi: https://doi.org/10.1038/s41598-019-43031-x (2019).

23 Palmer, D. S., O'Boyle, N. M., Glen, R. C. & Mitchell, J. B. O. Random Forest Models To Predict Aqueous Solubility. *Journal of Chemical Information and Modeling* **47**, 150-158, doi: https://doi.org/10.1021/ci060164k (2007).

24 Tao, P. *et al.* Solar-driven interfacial evaporation. *Nature Energy* **3**, 1031-1041, doi: https://doi.org/10.1038/s41560-018-0260-7 (2018).

25 Peng, G. *et al.* High efficient solar evaporation by airing multifunctional textile. *International Journal of Heat and Mass Transfer* **147**, 118866, doi:https://doi.org/10.1016/j.ijheatmasstransfer.2019.118866 (2020).

26 Chen, C., Kuang, Y. & Hu, L. Challenges and Opportunities for Solar Evaporation. *Joule* **3**, 683-718, doi:https://doi.org/10.1016/j.joule.2018.12.023 (2019).

27 Sharshir, S. W. *et al.* Influence of basin metals and novel wick-metal chips pad on the thermal performance of solar desalination process. *Journal of Cleaner Production* **248**, 119224, doi:https://doi.org/10.1016/j.jclepro.2019.119224 (2020).

28 Poós, T. & Varju, E. Review for prediction of evaporation rate at natural convection. *Heat and Mass Transfer* **55**, 1651-1660, doi: https://doi.org/10.1007/s00231-018-02535-4 (2019).

29 Wang, Y. *et al.* A New Machine Learning Algorithm to Optimize A Reduced Mechanism of 2-Butanone and the Comparison with Other Algorithms. *ES Materials & Manufacturing* **6**, 28-37, doi: https://doi.org/10.30919/esmm5f615 (2019).

30 Weber, D. G. G. a. R. L. S. a. H. K. M. a. B. W. Cantera: An Object-oriented Software Toolkit for Chemical Kinetics, Thermodynamics, and Transport Processes. doi: https://doi.org/10.5281/zenodo.4527812 (2021).